Further Examples of Evolution by Gene Duplication Revealed Through DNA Sequence Comparisons

Tomoko Ohta

National Institute of Genetics, Mishima 411, Japan

Manuscript received March 29, 1994
Accepted for publication August 22, 1994

ABSTRACT

To test the theory that evolution by gene duplication occurs as a result of positive Darwinian selection that accompanies the acceleration of mutant substitutions, DNA sequences of recent duplication were analyzed by estimating the numbers of synonymous and nonsynonymous substitutions. For the troponin C family, at the period of differentiation of the fast and slow isoforms, amino acid substitutions were shown to have been accelerated relative to synonymous substitutions. Comparison of the first exon of α-actin genes revealed that amino acid substitutions were accelerated when the smooth muscle, skeletal and cardiac isoforms differentiated. Analysis of members of the heat shock protein 70 gene family of mammals indicates that heat shock responsive genes including duplicated copies are evolving rapidly, contrary to the cognitive genes which have been evolutionarily conservative. For the α_1-antitrypsin reactive center, the acceleration of amino acid substitution has been found for gene pairs of recent duplication.

EVOLUTION by gene duplication has now been widely accepted as an important phenomenon (OHNO 1970; OHTA 1980). Based on theoretical study, I suggested that positive Darwinian selection is needed for acquiring a gene with a modified function, whereas genes whose function has been fixed for a long time evolve mostly through random genetic drift (OHTA 1988). There are now several examples that are compatible with the above theory, stomach lysozyme (IRWIN and WILSON 1990), visual pigment genes (YOKOYAMA and YOKOYAMA 1990), homeobox family (KAPPEN *et al.* 1989), hemoglobin γ (FITCH *et al.* 1991), alcohol dehydrogenase (LONG and LANGLEY 1993), ion channel family (STRONG *et al.* 1993) and growth hormone family (OHTA 1993; WALLIS 1993). There are other examples of accelerated amino acid substitution connected to duplication, in which the underlying mechanism is not obvious (LI 1985).

To examine the mechanisms of evolution by gene duplication, it is desirable to clarify the pattern of nucleotide substitutions of recently duplicated genes, because old differentiation is masked by subsequent nucleotide substitutions and therefore often undetectable. Especially if one wants to determine the relative numbers of synonymous and nonsynonymous substitutions among duplicated genes, the former may be saturated by old duplication, and a reliable estimate cannot be obtained.

What do the relative numbers of synonymous and nonsynonymous substitutions tell us? Because most of synonymous substitutions are thought to be selectively neutral in mammals, they accumulate by mutation rate (KIMURA 1983). If nonsynonymous substitutions show a different pattern, selection should be responsible for the pattern. It may be argued that the different pattern simply reflects change in selective constraint and does not show the positive selection. In fact, it is customary to suppose that positive selection is detectable only when the number of nonsynonymous substitutions exceeds that of synonymous substitutions (HUGHES and NEI 1988). Although the explanation by changing constraints cannot be completely denied, the operation of positive selection is a more natural interpretation, if the acceleration of amino acid substitutions is detected in conjunction with functional differentiation.

The data for nucleotide sequences in the gene families of troponin C, α-actin, heat shock protein and antitrypsin include those of some recently duplicated genes with distinct functions. The pattern of synonymous and nonsynonymous nucleotide substitutions of such sequences were examined. Nonsynonymous substitutions were shown to be accelerated in conjunction with functional differentiation.

DATA ANALYSIS

Nucleotide sequences were obtained from the genetic databases maintained at the National Institute of Genetics, which include GenBank, DDBJ (DNA Databank of Japan), and EMBL. Five sequences of troponin C, nine sequences of α-actin, 8 sequences of heat shock protein 70 and 10 sequences of α_1-antitrypsin were used. For the acquisition and analysis of the data, the ODEN Package created by INA (1992) was used. The numbers of synonymous and nonsynonymous substitutions were estimated using the method of NEI and GOJOBORI (1986) which is in the ODEN Package. This method divides nucleotide substitutions into synonymous and nonsynonymous categories, and then the multiple hit is estimated under the assumption of random mutability among the four kinds of bases. The standard error of the

TABLE 1

The number of synonymous substitutions (upper figures) and that of nonsynonymous substitutions (lower figures) per 100 sites among troponin C genes of human, mouse and rabbit

		Slow troponin C		Fast troponin C		
		MUSCTNCA	HUMTNCS	MUSFSTC6	RABTNC	HUMTC2
Slow	MUSCTNCA		70.37 ± 13.27 *40.62 ± 7.75*	236.29 ± 111.07 *189.51 ± 143.94*	108.81 ± 22.26 *111.86 ± 31.00*	107.35 ± 21.89 *109.53 ± 30.33*
	HUMTNCS	0.26 ± 0.26 *0.30 ± 0.30*		155.73 ± 39.75 *134.14 ± 46.92*	94.25 ± 18.35 *90.25 ± 20.84*	76.57 ± 14.55 *84.37 ± 18.45*
Fast	MUSFSTC6	26.10 ± 2.99 *27.78 ± 3.20*	25.35 ± 2.94 *26.55 ± 3.09*		70.80 ± 13.37 *42.27 ± 8.03*	77.11 ± 14.62 *43.38 ± 8.41*
	RABTNC	26.37 ± 3.02 *26.80 ± 3.07*	25.62 ± 2.96 *26.21 ± 3.04*	0.26 ± 0.26 *0.30 ± 0.30*		28.92 ± 6.56 *19.97 ± 4.52*
	HUMTC2	26.16 ± 3.00 *26.65 ± 3.06*	25.35 ± 2.94 *25.17 ± 2.92*	0.52 ± 0.37 *0.61 ± 0.43*	0.26 ± 0.26 *0.29 ± 0.29*	

The Roman figures were obtained using the method of NEI-GOJOBORI (1986) and the italic figures, with that of INA (1994). As for the accession number and the data source, see APPENDIX.

estimate was calculated using the method of KIMURA and OHTA (1972) which is also in the ODEN Package.

Since the assumption of random base substitution is often not satisfied, the method of NEI-GOJOBORI is not quite satisfactory. INA (1994) invented a new method which brings KIMURA's two-parameter model into the NEI-GOJOBORI method. Through extensive simulations, INA has shown that his method usually gives a better estimate than that of NEI-GOJOBORI, and his method has been used in this study. On the other hand, LI's (1993) new method, which gives a similar estimate to INA's, has turned out to be often inapplicable for the present data sets and is not used here. For the present purpose of finding the acceleration of nonsynonymous substitutions in conjunction with functional differentiation, an analysis using the two methods gives satisfactory results. In INA's method, the ratio of transition to transversion is estimated from the third position of codons, and the estimated value is used for calculating the numbers of synonymous and nonsynonymous substitutions. When the sequence used was short, *i.e.*, less than 150 nucleotide sites, the ratio was estimated beforehand, by using the larger region that includes the sequence.

RESULTS OF ANALYSES

Troponin C gene family: Troponin C is a protein that regulates excitation-contraction coupling in heart and skeletal muscle. In mammals, two distinct isoforms of this protein have been identified, the fast and the slow types, which are encoded by homologous but different genes (DHOOT and PERRY 1979). DNA sequences of both genes in mouse and man were available and analyzed.

Table 1 shows the results. The upper figures are the estimated number of synonymous substitutions with a standard error per 100 sites, and the lower figures, that of nonsynonymous substitutions. The Roman figures were calculated using the method of NEI-GOJOBORI, and the italic figures, with that of INA. Note that the estimated values by the two methods considerably differ in some sequence pairs, but are very similar in other pairs. This is because the difference of the estimated values depends on the magnitude of bias in transition-transversion ratio. For details of such effects, see INA (1994).

A striking pattern was found; the number of nonsynonymous substitutions relative to that of synonymous substitutions was much smaller for the slow-slow and for the fast-fast comparisons than for the slow-fast comparisons. In other words, amino acid substitution must have been accelerated during the period of differentiation of the slow and the fast troponins, and then was slowed down subsequently. According to PARMACEK *et al.* (1990), the divergence between the two forms of troponin C is associated with their functional differentiation. This statement has now been expanded so that functional differentiation is associated with rapid amino acid substitution. In this process, positive Darwinian selection is the most likely cause of the acceleration.

α-Actin genes: Actin is the most abundant of the cytoskeletal proteins. Actin genes form a highly conserved family in eucaryote genomes. Several isoforms of α-actin are known such as skeletal, cardiac and smooth muscle types (HU *et al.* 1986). By comparing sequences of these isoforms, it is found that about half of the nonsynonymous substitutions are within the first exon that is only about 12% of the protein. It has also been reported that the amino terminus (first exon) is the site of major actin-myosin interactions (SUTOH 1982). Thus, it is likely that the functional differentiation among the isoforms, if any, may be revealed by the first exon sequence. In the following, the result of sequence analysis of the first exon is presented.

In Table 2, the estimated numbers of synonymous and nonsynonymous substitutions are given as before. Again, the number of nonsynonymous substitutions relative to

TABLE 2

The number of synonymous substitutions (upper figures) and that of nonsynonymous substitutions (lower figures) per 100 sites of the first exon among actin genes of human, mouse, rat and rabbit

		Smooth muscle α-actin				Skeletal α-actin			Cardiac α-actin	
		RNACTAV	MMACTASM	HSACTA	OCRNAASMA	HUMACTASK	MUSACASA	RNAC 02	HUMACTCA1	MUSACTCA5
Smooth muscle	RNACTAV		22.67 ± 9.69 *18.55 ± 8.22*	49.55 ± 16.90 *39.44 ± 14.21*	71.01 ± 23.26 *55.38 ± 18.98*	224.48 ± 160.66 *147.15 ± 94.63*	100.68 ± 33.86 *70.77 ± 20.98*	80.05 ± 25.90 *58.49 ± 17.34*	146.73 ± 60.03 *101.63 ± 38.59*	81.20 ± 26.41 *59.91 ± 18.85*
	MMACTASM	0.00 *0.00*		37.76 ± 13.72 *30.46 ± 11.64*	101.95 ± 35.02 *86.73 ± 38.31*	150.43 ± 63.11 *108.70 ± 46.98*	150.43 ± 63.11 *103.26 ± 39.82*	114.61 ± 40.44 *80.96 ± 26.04*	111.83 ± 38.84 *81.57 ± 27.57*	133.78 ± 51.58 *104.58 ± 48.34*
	HSACTA	0.00 *0.00*	0.00 *0.00*		49.94 ± 17.06 *41.96 ± 16.09*	241.42 ± 200.80 —	116.81 ± 41.73 *88.33 ± 34.00*	117.57 ± 42.19 *89.27 ± 34.67*	151.72 ± 64.21 *107.87 ± 46.22*	119.90 ± 43.61 *90.60 ± 36.51*
	OCRNAASMA	0.00 *0.00*	0.00 *0.00*	1.02 ± 1.02 *1.10 ± 1.10*		232.43 ± 178.38 —	101.82 ± 34.44 *76.91 ± 26.43*	102.39 ± 34.74 *77.61 ± 26.82*	149.17 ± 62.05 *99.80 ± 35.16*	104.17 ± 35.68 *75.78 ± 25.26*
Skeletal	HUMACTASK	6.99 ± 2.78 *7.59 ± 3.04*	6.99 ± 2.78 *7.59 ± 3.04*	8.09 ± 3.00 *8.80 ± 3.28*	6.98 ± 2.70 *7.56 ± 3.03*		57.16 ± 18.41 *50.74 ± 19.06*	80.49 ± 25.68 *77.34 ± 34.04*	79.04 ± 25.06 *60.65 ± 18.71*	81.62 ± 26.18 *63.57 ± 21.02*
	MUSACASA	8.12 ± 3.01 *8.84 ± 3.31*	8.12 ± 3.01 *8.84 ± 3.31*	9.24 ± 3.22 *10.01 ± 3.54*	8.11 ± 3.00 *8.80 ± 3.30*	0.00 *0.00*		13.48 ± 6.91 *11.58 ± 6.17*	88.35 ± 28.36 *65.40 ± 19.28*	91.47 ± 29.80 *71.68 ± 24.81*
	RNAC 02	8.12 ± 3.01 *8.82 ± 3.30*	8.12 ± 3.01 *8.82 ± 3.30*	9.23 ± 3.22 *10.04 ± 3.54*	8.10 ± 3.00 *8.78 ± 3.29*	0.00 *0.00*	0.00 *0.00*		88.78 ± 28.55 *65.88 ± 19.47*	73.24 ± 23.39 *57.01 ± 18.38*
Cardiac	HUMACTCA1	8.15 ± 3.02 *8.87 ± 3.32*	8.15 ± 3.02 *8.87 ± 3.32*	8.12 ± 3.01 *8.85 ± 3.31*	8.13 ± 3.01 *8.83 ± 3.31*	3.18 ± 1.85 *3.43 ± 2.00*	3.18 ± 1.85 *3.43 ± 2.00*	3.18 ± 1.85 *3.42 ± 1.99*		45.17 ± 15.19 *37.42 ± 13.31*
	MUSACTA5	8.09 ± 3.00 *8.84 ± 3.31*	8.09 ± 3.00 *8.84 ± 3.31*	8.07 ± 2.99 *8.82 ± 3.30*	9.23 ± 3.22 *10.04 ± 3.53*	3.16 ± 1.84 *3.41 ± 1.98*	3.16 ± 1.84 *3.41 ± 1.98*	3.16 ± 1.84 *3.41 ± 1.98*	1.04 ± 1.04 *1.12 ± 1.12*	

See Table 1. Dashes indicate unestimable cases.

TABLE 3

The number of synonymous substitutions (upper figures) and that of nonsynonymous substitutions (lower figures) per 100 sites among gene members of HSP70 family

	Heat shock responsive			Heat shock cognate			Testis specific	
	HUMP70B	MUSHSP7A2	HUMHSP70D	BOVHSCP	HSHSC70	MUSHSPCA	RATHST70A	MUSHSC70B
HUMP70B		72.30 ± 6.17 *67.89 ± 5.90*	70.17 ± 6.02 *61.75 ± 5.46*	239.03 ± 52.23 *163.66 ± 35.32*	— *210.04 ± 86.06*	166.23 ± 21.08 *129.32 ± 18.66*	106.78 ± 10.00 *92.83 ± 9.35*	101.93 ± 9.23 *89.75 ± 8.74*
MUSHSP7A2	11.08 ± 0.94 *11.28 ± 0.96*		32.84 ± 3.25 *34.39 ± 3.42*	— —	— —	— *218.74 ± 101.67*	91.74 ± 8.27 *89.84 ± 8.40*	89.90 ± 7.92 *85.41 ± 7.79*
HUMHSP70D	10.57 ± 0.92 *10.98 ± 0.96*	2.48 ± 0.43 *2.45 ± 0.42*		— —	— —	— *189.48 ± 57.38*	85.71 ± 7.68 *79.04 ± 7.36*	84.56 ± 7.42 *79.42 ± 7.20*
BOVHSCP	13.82 ± 1.07 *14.84 ± 1.15*	10.26 ± 0.90 *11.22 ± 0.99*	9.53 ± 0.87 *10.45 ± 0.95*		46.66 ± 4.31 *34.69 ± 3.34*	74.71 ± 6.69 *54.04 ± 5.16*	— —	— —
HSHSC70	13.66 ± 1.06 *14.75 ± 1.15*	10.18 ± 0.90 *11.24 ± 0.99*	9.36 ± 0.86 *10.38 ± 0.95*	0.14 ± 0.10 *0.16 ± 0.11*		68.60 ± 6.14 *50.41 ± 4.82*	— —	— —
MUSHSPCA	13.66 ± 1.08 *14.52 ± 1.15*	10.31 ± 0.92 *11.10 ± 0.99*	9.42 ± 0.88 *10.14 ± 0.94*	0.37 ± 0.16 *0.40 ± 0.18*	0.36 ± 0.16 *0.40 ± 0.18*		— *251.38 ± 200.06*	— —
RATHST70A	14.37 ± 1.11 *14.93 ± 1.15*	11.13 ± 0.96 *11.21 ± 0.97*	11.32 ± 0.97 *11.57 ± 0.99*	8.57 ± 0.83 *9.51 ± 0.92*	8.62 ± 0.83 *9.44 ± 0.91*	8.88 ± 0.86 *9.53 ± 0.92*		13.82 ± 1.96 *11.51 ± 1.64*
MUSHSC70B	14.31 ± 1.09 *14.83 ± 1.13*	10.89 ± 0.93 *11.05 ± 0.95*	11.20 ± 0.95 *11.39 ± 0.97*	8.66 ± 0.82 *9.55 ± 0.90*	8.71 ± 0.82 *9.57 ± 0.91*	8.77 ± 0.84 *9.44 ± 0.90*	0.40 ± 0.17 *0.43 ± 0.18*	

See Table 1. Dashes indicate unestimable cases.

that of synonymous substitutions is smaller for the same-isoform comparisons than for the different-isoform comparisons. In fact, for most of the within-isoform comparisons, no nonsynonymous substitution is observed. Both of two exceptions, HSACT-OCRNAASMA pair of the smooth muscle actin and HUMACTCA1-MUSACTCA5 pair of the cardiac actin, are caused by a two-step change of a codon, and there is no amino acid difference. Note that, in counting the numbers of synonymous and nonsynonymous substitutions, all possible paths are equally weighted for multiple-step changes of a codon in the NEI-GOJOBORI and the INA methods. Thus, it is likely that there is no amino acid substitution among the same-isoform sequences. As for the different-isoform comparisons, there are 3–8.5 nonsynonymous differences that result in amino acid differences. Actually, the skeletal actin and the cardiac actin are coexpressed and their tissue specificity is not complete (GUNNING *et al.* 1983). The amino acid difference between the two isoforms is smaller than that between the smooth muscle actin and skeletal or cardiac actin. The amino acid divergence seems to correlate well with the tissue specificity. The estimated numbers of synonymous and nonsynonymous substitutions have large standard errors because of the short region counted, and statistical significance can not be obtained for the differences of the ratio of the nonsynonymous to the synonymous substitutions. However, the tendency is clear, and it is likely that some amino acid substitutions have been caused by the functional differentiation of binding with the tissue specific myosin.

Heat shock protein 70 gene family: Heat shock induces several kinds of proteins in most organisms, among which heat shock protein (hsp) 70 is the foremost. In mammals, this protein is encoded by a multigene family (LINDQUIST 1986), and several DNA sequences of mammals are available. The multigene family includes heat shock responsive and nonresponsive genes, and the latter participates in regulation of ordinary cell growth (PELHAM 1986; GIEBEL *et al.* 1988).

DNA sequences of hsp70 genes of several mammalian species were analyzed; three sequences of the heat shock responsive gene, three sequences of the heat shock cognate gene and two sequences of the testis-specific gene. Table 3 presents the results. As before, the upper figures are the estimated number with standard error of synonymous substitutions and the lower figures, that of nonsynonymous substitutions per 100 sites. A dash indicates the inestimable case. Again, the relative values of nonsynonymous and synonymous substitution numbers show an interesting pattern. Namely, the number of nonsynonymous substitutions relative to that of synonymous ones is quite small for comparisons within the cognate or within the testis-specific group, but it is not so for comparisons among heat-shock-inducible genes and for comparisons between heat-shock-inducible and testis-specific genes. The cognate group appears to be too divergent from the other groups to get a reliable estimate of the number of synonymous substitutions.

Sequence divergence within the heat-shock-inducible group needs more detailed examination. Note that the two human sequences of this group are the product of gene duplication which occurred before the human-mouse divergence, and HUMHSP70D is orthologous to the mouse gene. It can be seen that the number of nonsynonymous substitutions relative to that of synonymous ones is smaller between the human-mouse orthologous comparison than

TABLE 4

The number of synonymous substitutions (upper figures) and that of nonsynonymous substitutions (lower figures) per 100 sites of reactive center among gene members of α_1-antitrypsin family

	MUSA1AT	MUSAIPI1A	MUSAIPI4A	MUSAIPI5A	MUSAIPI2A	MUSAIPI3A	RATATRA1	GPIAPSA1	RABA1APTF	HUMA1ATZ
MUSA1AT		4.53 ± 5.59 *3.97 ± 4.91*	0.00 *0.00*	10.48 ± 8.72 *8.91 ± 7.46*	0.00 *0.00*	6.32 ± 6.39 *5.72 ± 5.89*	74.89 ± 34.06 *68.11 ± 35.60*	151.57 ± 96.73 *99.94 ± 46.69*	— —	172.12 ± 122.82 *194.30 ± 422.91*
MUSAIPI1A	19.83 ± 7.14 *21.19 ± 7.80*		3.23 ± 4.59 *2.89 ± 4.11*	10.61 ± 8.83 *9.07 ± 7.60*	4.48 ± 5.53 *3.99 ± 4.94*	10.97 ± 8.67 *9.82 ± 7.80*	100.34 ± 47.87 *88.40 ± 47.75*	113.73 ± 60.71 *90.38 ± 51.00*	— —	— —
MUSAIPI4A	31.21 ± 9.64 *32.95 ± 10.23*	20.61 ± 7.37 *21.60 ± 7.74*		0.00 *0.00*	0.00 *0.00*	6.06 ± 6.13 *5.59 ± 5.76*	77.98 ± 34.80 *71.89 ± 36.51*	104.26 ± 52.30 *85.50 ± 45.30*	161.70 ± 102.58 *137.96 ± 105.73*	— —
MUSAIPI5A	20.36 ± 7.27 *21.76 ± 7.80*	25.97 ± 8.46 *28.61 ± 9.78*	25.02 ± 8.33 *26.67 ± 8.94*		6.75 ± 6.83 *5.92 ± 6.11*	6.32 ± 6.39 *5.59 ± 5.76*	74.89 ± 34.06 *65.43 ± 33.46*	185.47 ± 147.86 *130.54 ± 105.24*	179.30 ± 130.66 *132.22 ± 96.93*	161.38 ± 107.39 *152.54 ± 187.04*
MUSAIPI2A	16.41 ± 6.40 *17.22 ± 6.74*	12.12 ± 5.37 *12.73 ± 5.68*	22.20 ± 7.74 *23.27 ± 8.17*	24.71 ± 8.22 *26.48 ± 8.95*		6.25 ± 6.32 *5.75 ± 5.93*	73.56 ± 33.27 *64.42 ± 30.19*	93.96 ± 46.65 *76.02 ± 38.49*	152.77 ± 93.25 *124.59 ± 80.80*	— —
MUSAIPI3A	25.34 ± 8.45 *26.85 ± 9.05*	20.40 ± 7.36 *21.38 ± 7.73*	9.28 ± 4.72 *9.69 ± 4.94*	28.42 ± 9.10 *30.47 ± 9.88*	16.87 ± 6.59 *17.95 ± 7.20*		91.78 ± 41.16 *89.55 ± 51.99*	120.71 ± 63.25 *105.65 ± 68.70*	— —	— —
RATATRA1	29.49 ± 9.26 *31.90 ± 0.44*	35.32 ± 10.47 *37.29 ± 11.13*	42.17 ± 12.03 *45.01 ± 13.22*	47.11 ± 13.01 *51.16 ± 14.51*	23.55 ± 8.02 *25.01 ± 8.71*	35.47 ± 10.65 *38.62 ± 12.35*		89.72 ± 43.64 *78.88 ± 45.29*	73.39 ± 33.54 *61.07 ± 26.81*	108.31 ± 54.43 *86.30 ± 41.88*
GPIAPSA1	42.67 ± 11.80 *45.86 ± 12.85*	58.33 ± 15.21 *62.85 ± 16.67*	79.86 ± 20.82 *87.76 ± 24.32*	70.65 ± 18.23 *78.30 ± 20.86*	52.50 ± 13.95 *56.16 ± 15.12*	82.40 ± 21.65 *90.94 ± 25.62*	45.12 ± 12.38 *48.24 ± 13.39*		98.68 ± 50.17 *77.81 ± 39.91*	97.45 ± 50.96 *70.68 ± 32.75*
RABA1APTF	52.25 ± 14.06 *55.90 ± 15.26*	58.51 ± 15.47 *64.40 ± 18.57*	81.87 ± 21.69 *90.69 ± 27.00*	80.84 ± 21.24 *92.12 ± 27.70*	56.86 ± 15.14 *61.48 ± 17.27*	83.48 ± 22.28 *92.80 ± 28.08*	51.63 ± 13.99 *55.06 ± 15.20*	34.04 ± 9.99 *36.24 ± 10.71*		109.78 ± 56.74 *101.05 ± 69.88*
HUMA1ATZ	27.88 ± 8.79 *29.56 ± 9.36*	55.57 ± 14.66 *64.40 ± 18.57*	64.83 ± 16.98 *69.52 ± 18.59*	61.85 ± 16.15 *67.67 ± 18.08*	42.86 ± 11.92 *45.39 ± 12.73*	56.41 ± 15.06 *60.32 ± 16.39*	25.76 ± 8.39 *27.65 ± 8.89*	45.09 ± 12.19 *48.96 ± 13.71*	50.05 ± 13.43 *54.25 ± 15.20*	

See Table 1. Dashes indicate unestimable cases.

between paralogous comparisons. The difference in relative values is statistically significant. Thus the general pattern of the rapid amino acid divergence in connection with functional differentiation is observed for the hsp70 gene family.

α_1-Antitrypsin gene family: Proteases and their inhibitors are encoded by gene families in mammals. Fortunately, some of their tertiary structures have been analyzed, and the regions of the reactive center have been determined. Because of the hypervariability of amino acids at the reactive center, it has been proposed that positive natural selection has operated to increase the rate of amino acid substitutions (LASKOWSKI *et al.* 1987; HILL and HASTIE 1987). Gene conversion has also been suggested as promoting variability (OHTA and BASTEN 1992). Here it is desirable to examine how this hypervariability is related to gene duplication. Several mouse gene sequences of α_1-antitrypsin available from GenBank, apparently have occurred by recent duplication, and show hypervariability at the reactive center (BORRIELLO and KRAUTER 1991). In this report, mouse sequences were analyzed together with sequences from other species; human, rabbit, guinea pig and rat (see APPENDIX). The results for the reactive center region are presented, which include 21 amino acid sites.

Table 4 shows the results. The upper and lower figures are the numbers of synonymous and nonsynonymous substitutions as before. From the table, it can be seen that, among the mouse sequences, the nonsynonymous substitution number exceeds the synonymous substitution number, but that, for sequence pairs between species, the former does not exceed the latter. It is likely that the acceleration of amino acid substitution occurred through selection for mouse genes in conjunction with gene duplication.

DISCUSSION

The present analyses provide more examples of evolution by gene duplication in which functional differentiation accompanies acceleration of amino acid substitution (OHTA 1991). Although only coding regions were analyzed, these examples also suggest that evolution of regulatory elements is important, since functional differentiation is correlated with changes of expression. It would be desirable to study the differentiation pattern of regulatory elements in the future. Actually, one would expect that more duplicated genes deteriorate than observed from the theoretical point of view (KIMURA and KING 1979; OHTA 1988). One reason for less chance of their deterioration than expected would be the prevention of deterioration through continued expression. Unless regulatory elements change, abnormal products would be harmful and would be selected against. Then protein sequences and regulatory elements would have to coevolve. In such a situation, acceleration of amino acid substitution is likely to be caused by positive Darwinian selection rather than by simple relaxation of selective constraints.

In eukaryote genomes, particularly in mammalian genomes, there are numerous gene families that originated mostly by gene duplication. However, in many cases, the duplication event was ancient, and synonymous substitutions are saturated. Then the present approach is not applicable. It should be noted that, even in the present results, the synonymous substitution number per 100 sites is not very reliable, when the number is 100 or more in the tables. The purpose of this study is to show the pattern of synonymous *vs.* nonsynonymous substitutions in duplicated genes, and not to estimate the divergence accurately.

Another related problem in the present study is the concept of a molecular clock. It is now clear that the rate of amino acid substitution varies according to the functional differentiation of the products, which is often associated with gene duplication. Amino acid sequences are commonly used for constructing phylogenetic trees. One has to be careful in interpreting the results because of the acceleration of amino acid substitution at the period of functional differentiation.

I thank WEN-HSIUNG LI and an anonymous referee for their valuable comments on the manuscript. This work was supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan. This is contribution no. 1999 from the National Institute of Genetics, Mishima 411, Japan.

LITERATURE CITED

BORRIELLO, F., and K. S. KRAUTER, 1991 Multiple murine α_1-protease inhibitor genes show unusual evolutionary divergence. Proc. Natl. Acad. Sci. USA **88:** 9417–9421.

DHOOT, G. K., and S. V. PERRY, 1979 Distribution of polymorphic forms of troponin components and tropomyosin in skeletal muscle. Nature **278:** 714–718.

FITCH, D. H. A., W. J. BAILEY, D. A. TAGLE, M. GOODMAN, L. SIEU *et al.*, 1991 Duplication of the γ-globin gene mediated by repetitive L1 LINE sequences in an early ancestor of simian primates. Proc. Natl. Acad. Sci. USA **88:** 7396–7400.

GIEBEL, L. B., B. P. DWORNICZAK and E. K. F. BAUTZ, 1988 Developmental regulation of a constitutively expressed mouse mRNA encoding a 72-kDa heat shock-like protein. Dev. Biol. **125:** 200–207.

GUNNING, P., P. PONTE, H. BLAU and L. KEDES, 1983 α-skeletal and α-cardiac actin genes are coexpressed in adult human skeletal muscle and heart. Mol. Cell. Biol. **3:** 1985–1995.

HILL, R. E., and N. D. HASTIE, 1987 Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature **326:** 96–99.

HU, M. C.-T., S. B. SHARP and N. DAVIDSON, 1986 The complete sequence of the mouse skeletal α-actin gene reveals several conserved and inverted repeat sequence outside of the protein-coding region. Mol. Cell. Biol. **6:** 15–25.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature **335:** 167–170.

INA, Y., 1992 *ODEN.* National Institute of Genetics, Mishima, Japan.

INA, Y., 1994 New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. (in press).

IRWIN, D. M., and A. C. WILSON, 1990 Concerted evolution of ruminant stomach lysozymes. J. Biol. Chem. **265:** 4944–4952.

KAPPEN, C., K. SCHUGHART and F. H. RUDDLE, 1989 Two steps in the evolution of antennapedia-class vertebrate homeobox genes. Proc. Natl. Acad. Sci. USA **86:** 5459–5463.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KIMURA, M., and J. L. KING, 1979 Fixation of a deleterious allele at one of two *duplicate* loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA **76:** 2858–2861.

KIMURA M., and T. OHTA, 1972 On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2:** 87–90.

LASKOWSKI, M., JR., I. KATO, W. J. KOHR, S. J. PARK, M. TASHIRO *et al.*, 1987 Positive Darwinian selection in evolution of protein inhibitors of serine proteinases. Cold Spring Harbor Symp. Quant. Biol. **52:** 545–553.

LI, W.-H., 1985 Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy, pp. 333–352 in *Population Genetics and Molecular Evolution*, edited by T. OHTA and K. AOKI. Japan Scientific Societies Press, Tokyo.

LI, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

LINDQUIST, S., 1986 The heat shock response. Annu. Rev. Biochem. **55:** 1151–1191.

LONG, M., and C. H. LANGLEY, 1993 Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science **260:** 91–95.

NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

OHNO, S., 1970 *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

OHTA, T., 1980 *Evolution and Variation of Multigene Families* (Lecture. Notes in Biomathematics, Vol. 37). Springer-Verlag, Berlin.

OHTA, T., 1988 Further simulation studies on evolution by gene duplication. Evolution **42:** 375–386.

OHTA, T., 1991 Multigene families and the evolution of complexity. J. Mol. Evol. **33:** 34–41.

OHTA, T., 1993 Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134:** 1271–1276.

OHTA, T., and C. J. BASTEN, 1992 Gene conversion generates hypervariability at the variable regions of kallikreins and their inhibitors. Mol. Phyl. Evol. **1:** 87–90.

PARMACEK, M. S., A. R. BENGUR, A. J. VARA and J. M. LEIDEN, 1990 The structure and regulation of expression of the murine fast skeletal troponin C gene. J. Biol. Chem. **265:** 15970–15976.

PELHAM, H. R. B., 1986 Speculations on the functions of the major heat shock and glucose-regulated proteins. Cell **46:** 959–961.

STRONG, M., K. G. CHANDY and G. A. GUTMAN, 1993 Molecular evolution of voltage-sensitive ion channel genes: on the origin of electrical excitability. Mol. Biol. Evol. **10:** 221–242.

SUTOH, K., 1982 Identification of myosin binding sites on the actin sequence. Biochemistry **21:** 3654–3661.

WALLIS, M., 1993 Remarkably high rate of molecular evolution of ruminant placental lactogens. J. Mol. Evol. **37:** 86–88.

YOKOYAMA, S., and R. YOKOYAMA, 1990 Molecular evolution of visual pigment genes and other G-protein-coupled genes, pp. 307–322 in *Population Biology of Genes and Molecules*, edited by N. TAKAHATA and J. F. CROW,. Baifukan, Tokyo.

Communicating editor: W-H. LI

APPENDIX

Sequences from mouse, human, rabbit, guinea pig and rat used are given in Table 5.

TABLE 5

Accession number and data source

	Accession no.	Data source
Troponin C		
MUSCTNCA	M29793	PARMACEK and LEIDEN, *J. Biol. Chem.* **264,** 13217 (1989)
HUMTNCS	X07897	GAHLMANN *et al.*, *J. Mol. Biol.* **201,** 379 (1988)
MUSFSTC6	M57590	PARMACEK *et al.*, *J. Biol. Chem.* **265,** 15970 (1990)
RABTNC	J03462	ZOT *et al.*, *J. Biol. Chem.* **262,** 15418 (1987)
HUMTC2	X07898	GAHLMANN *et al.*, *J. Mol. Biol.* **201,** 379 (1988)
α-Actin		
RNACTAV	X06801	MCHUGH and LESSARD, *Nucleic Acids Res.* **16,** 4167 (1988)
MMACTASM	X13297	MIN *et al.*, *Nucleic Acids Res.* **16,** 10374 (1988)
HSACTA	X13839	KAMADA and KAKUNAGA, *Nucleic Acids Res.* **17,** 1767 (1989)
OCRNAASMA	X60732	HARRIS *et al.*, *Gene* **112,** 265 (1992)
HUMACTASK	J00068	GUNNING *et al.*, *Mol. Cell Biol.* **3,** 787 (1983)
MUSACASA	M12347	HU *et al.*, *Mol. Cell. Biol.* **6,** 15 (1986)
RNAC02	V01218	ZAKUT *et al.*, *Nature* **298,** 857 (1982)
HUMACTCA1	J00070	HAMADA *et al.*, *Proc. Natl. Acad. Sci. USA* **79,** 5901 (1982)
MUSACTA5	M59867	GARNER *et al.*, *EMBO J.* **5:** 2559 (1986)
hsp-70		
HUMP70B	X51757	LEUNG *et al.*, *Biochem. J.* **267,** 125 (1990)
MUSHSP7A2	M35021	HUNT and CALDERWOOD, *Gene* **87,** 199 (1990)
HUMHSP70D	M11717	HUNT and MORIMOTO, *Proc. Natl. Acad. Sci. USA* **82,** 6455 (1985)
BOVHSCP	X53827	DELUCA-FLAHERTY and MCKAY, *Nucleic Acids Res.* **18,** 5569 (1990)
HSHSC70	Y00371	DWORNICZAK and MIRAULT, *Nucleic Acids. Res.* **15,** 5181 (1987)
MUSHSPCA	M19141	GIEBEL *et al.*, *Dev. Biol.* **125,** 200 (1988)
RATHST70A	X15705	WISNIEWSKI *et al.*, *Biochim. Biophys. Acta* **1048,** 93 (1990)
MUSHSC70B	M20567	ZAKERI *et al.*, *Mol. Cell Biol.* **8,** 2925 (1988)
α_1-Antitrypsin		
MUSA1AT	M33567	LATIMER *et al.*, *Mol. Cell. Biol.* **10,** 760 (1990)
MSAIPI1A	M75721	BORRIELLO and KRAUTER, *Proc. Natl. Acad. Sci. USA* **88,** 9417 (1991)
MUSAIPI4A	M75718	BORRIELLO and KRAUTER, *Proc. Natl. Acad. Sci. USA* **88,** 9417 (1991)
MUSAIPI5A	M75717	BORRIELLO and KRAUTER, *Proc. Natl. Acad. Sci. USA* **88,** 9417 (1991)
MUSAIPI2A	M75716	BORRIELLO and KRAUTER, *Proc. Natl. Acad. Sci. USA* **88,** 9417 (1991)
MUSAIPI3A	M75720	BORRIELLO and KRAUTER, *Proc. Natl. Acad. Sci. USA* **88,** 9417 (1991)
RATATRA1	M32247	CHAO *et al.*, *Biochemistry* **29,** 323 (1990)
GPIAPSA1	M57270	SUZUKI *et al.*, *J. Biol. Chem.* **266,** 928 (1991)
RABA1APTF	D00853	SINOHARA and SAITO, *J. Biochem.* **109,** 158 (1991)
HUMA1ATZ	J02619	NUKIWA *et al.*, *J. Biol. Chem.* **261,** 15989 (1986)